QANet with Character Embedding

Gurdeep Sullan   Takara Truong

Introduction

While models such as BiDAF [1] show decent results on question-answer datasets like SQuAD, the reliance on recurrent structures makes both training and inference slow.

The slow and computationally expensive training time also makes a larger impact on the carbon footprint of these models. Compounded, these issues prevent such models from being deployed in industry.

In this project, we aim to find low cost and carbon footprint friendly improvements while achieving better performance than the BiDAF baseline on SQuAD.

Methods

Two improvements are added to the BiDaF model which is used as a baseline for experiments:

1. Character Level Embedding: Allows us to encode new or rare words in the input through character level compositions. This helps improve reading comprehension of the BiDaF model.
2. QANet: Uses transformer based encoder block which allows the model to hone in on important words within the context based on the query [2]. Unlike RNN based models which are prone to forgetting over long sequences and not parallelizable, attention mechanisms are able to access all hidden states in the sequence and are parallelizable.

Experiments

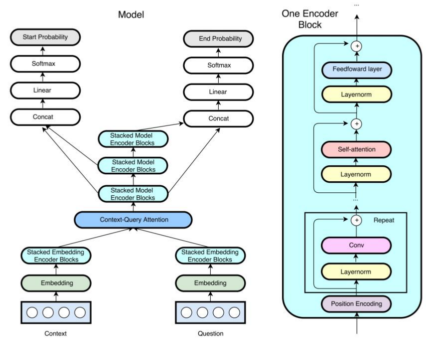
We performed an ablation study to understand how model architecture features influence performance. The model variants include:

- BiDAF
- QANet
- BiDAF + character embedding
- QANet + character embedding

Model

The general structure of the QANet model is an embedding layer, an embedding encoder block, a context-query attention layer, a series of model encoder blocks, and a final output layer.

The key repeated structure here is the encoder block, which comprises a positional encoder, a series of convolutional layers, a self attention layer, and a feedforward layer.



Results

- The QANet + character embedding model has the best performance on the devset when compared to other variants, Table 1. The result of this model on the test set: EM = 60.592, F1 = 64.353.
- Adding character embedding provides improvement in EM and F1 score on both models and is the more eco-friendly option than changing the architecture.
- The carbon footprint of training QANet + character embedding is equivalent to driving an average car for 8.37 km compared to BiDAF with character embedding at 1.29 km.

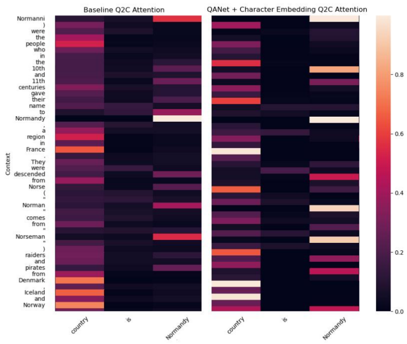
Table 1. Model Performance

Model	EM	F1	Carbon Footprint [kgC02eq]
BiDAF	61.32	58.13	0.32
QANet	50.78	53.35	2.08
BiDAF + char-embed	64.82	61.65	0.35
QANet + char-embed	**67.75**	**64.06**	2.09

Analysis

We find that QANet with character embedding is better able to attend to words in the context when compared to the BiDAF model as noted by larger distribution weights on key words.

Original query: In what country is Normandy located ?



Conclusion

Overall, we present character embedding and QANet as solutions that give decent improvements to the performance over the baseline model.

Future work to extend this model would look at applying the stochastic layer dropout as mentioned in the paper for improved generalizability. An extension to the QANet model beyond the paper would be to try it with Transformers-XL to improve longer-term dependencies

Stanford University

1. Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." arXiv preprint arXiv:1611.01603 (2016).
2. Yu, Adams Wei, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." arXiv preprint arXiv:1804.09541 (2018).

Stanford ENGINEERING
Computer Science

Gurdeep Sullan
Stanford University
Computer Science
gksullan@stanford.edu

Takara Truong
Stanford University
Computer Science
takaraet@stanford.edu